APPENDIX A: LIST OF COMMENTERS

	Commenter	Abbreviation
1.	@Link Networks	(@Link)
2.	American Association of Retired Persons	(AARP)
3.	AT&T Corporation	(AT&T)
4.	Adelphia Business Solutions	(Adelphia)
5.	Allegiance Telecom, Inc.	(Allegiance)
6.	Alliance for Public Technology	(APT)
7.	American Council of the Blind, American Foundation for the Blind, National Association of the Deaf, Telecommunications for the Deaf, Inc., and World Institute on Disability	(ACB)
8.	Association for Local Telecommunications Services	(ALTS)
9.	Cable and Wireless, Inc.	(C&W USA)
10.	Cablevision Lightpath, Inc.	(Lightpath)
11.	Choice One Communications, Inc.	(Choice One)
12.	City of New York	
13.	Closecall America, Inc.	(Closecall)
14.	Coalition to Ensure Responsible Billing	(CERB)
15.	Competition Policy Institute	(CPI)
16.	Competitive Telecommunications Association	(CompTel)
17.	Consortium for School Networking	(CoSN)
18.	Consumer Federation of America	(CFA)
19.	CoreComm Limited and CoreComm New York, Inc.	(CoreComm)
20.	Covad Communications Company	(Covad)
21.	Destek Networking Group, Inc.	(Destek)
22.	DSL.net, Inc.	(DSL.net)
23.	E.Spire Communications, Inc. & Net 2000 Communications Services, Inc.	(E.Spire)
24.	Excel Communications, Inc.	(Excel)
25.	Focal Communications Corporation of New York	(Focal-NY)
26.	General Services Administration	(GSA)
27.	Global NAPS, Inc.	
28.	ICG Telecom Group, Inc.	(ICG)
29.	Intermedia Communications, Inc.	(Intermedia)
30.	Keefe, Barbara, MainePOINT Project Director University of Maine System	
31.	Keep America Connected et. al	
32.	KMC Telecom, Inc.	(KMC)
33.	League of United Latin American Citizens, Brent Wilkes	(LULAC)
34.	MCI WorldCom, Inc.	(MCI)
35.	National ALEC Association	(NALA)

36.	National Association of Partners in Education	(NAPE)
37.	National Black Chamber of Commerce	(NBCC)
38.	National Consumers League	(NCL)
39.	National Small Business United	(NSBU)
40.	Nextlink New York, Inc.	(NEXTLINK)
41.	Network Access Solutions	(NAS)
42.	New England Conference of Public Utilities Commissioners	(NECPUC)
43.	NorthPoint Communications, Inc.	(NorthPoint)
44.	Ntegrity Telecontent Services, Inc.	(Ntegrity)
45.	Omnipoint Communications, Inc.	(Omnipoint)
46.	Organization of Chinese Americans, Inc.	(OCA)
47.	Organizations Concerned about Rural Education	(OCRE)
48.	Prism Communication Services, Inc.	(Prism)
49.	RCN Telecom Services, Inc.	(RCN)
50.	Rhythms Netconnections, Inc.	(Rhythms)
51.	Santo, Virginia	
52.	Sprint Communications Company, L.P.	(Sprint)
53.	State of New York Attorney General, Eliot Spitzer	
54.	New York Public Service Commission	(NYPSC)
55.	Telecommunications Resellers Association	(TRA)
56.	Teligent, Inc.	(Teligent)
57.	United Seniors Health Cooperative	(USHC)
58.	Z-Tel Communications, Inc.	(Z-Tel)

REPLY COMMENTS

	Commenter	Abbreviation
1.	AT&T Corporation	(AT&T)
2.	Allegiance Telecom, Inc.	(Allegiance)
3.	Association for Local Telecommunications Services	(ALTS)
4.	BellSouth	
5.	Communications Workers of America	(CWA)
6.	Competition Policy Institute	(CPI)
7.	Conversent Communications, LLC	(Conversent)
8.	Covad Communications Company	(Covad)
9.	DSL.net, Inc.	(DSL.net)
10.	Focal Communications Corporation of New York	(Focal)
11.	Keep America Connected et. al	
12.	Level 3 Communications, LLC	(Level 3)
13.	MCI WorldCom, Inc.	(MCI)
14.	MediaOne Group, Inc.	(MediaOne)
15.	National Association of Partners in Education	(NAPG)
16.	National Council on the Aging	(NCOA)
17.	National Education Association of New York	(NEA/NY)
18.	Network Access Solutions	(NAS)
19.	New York Public Service Commission	(NYPSC)
20.	NorthPoint Communications, Inc.	(NorthPoint)
21.	OmniPoint Communications, Inc.	(OmniPoint)
22.	Prism Communication Services, Inc.	(Prism)
23.	RCN Telecom Services, Inc.	(RCN)
24.	Rhythms Netconnections, Inc.	(Rhythms)
25.	State of New York Attorney General, Eliot Spitzer	
26.	Teligent, Inc.	(Teligent)
27.	U S WEST Communications, Inc.	(US WEST)

APPENDIX B: STATISTICAL METHODOLOGY

1. In this appendix, we discuss the statistical methodology and test statistics that Bell Atlantic employed in its application. We find that the modified z-test that Bell Atlantic uses for measurements with large sample sizes is an appropriate test. We also find that the tests that Bell Atlantic uses for measurements with small sample sizes, the binomial and t-tests, and the permutation tests, are also appropriate tests. We note that, in so concluding, we do not preclude the use of other statistical tests that have been developed in collaborative proceedings in other states. Finally, we discuss how we will use the z-scores provided in the Carrier to Carrier reports to determine if a difference in performance is statistically significant. We conclude that a 95 percent confidence level is the appropriate threshold to use for a determination of statistical significance.

2. When making a parity comparison, statistical analysis is a useful tool to take into account random variation in the metrics.[1] We note that random variation is inherent in the incumbent LEC's process of providing interconnection and access to unbundled network elements. Our concern is primarily that the process that the incumbent LEC employs be nondiscriminatory. Thus, the incumbent LEC could have a provisioning process that is identical in its ability to provide the same function to retail customers and to competitive LECs, but because of random factors outside the control of the BOC, the average completed interval could vary for retail customers and competitive LECs from month to month, such that for one particular month, the metric for competitors would show a longer average interval than would the metric for Bell Atlantic's retail customers. Thus, metric results showing weaker performance to competitors could be due to random variation in the measures, even though the process is inherently nondiscriminatory. Therefore, the use of statistical analysis to take into account random variation in the metrics is desirable.[2]

3. Statistical tests can be used as a tool in determining whether a difference in the measured values of two metrics means that the metrics probably measure two different processes, or instead that the two measurements are likely to have been produced by the same process. This can be done using traditional hypothesis testing.[3] Hypothesis testing involves testing to determine which of two hypotheses, usually called the null and the alternative hypotheses, is likely to be correct.[4] Usually this means devising a statistical test to determine whether the null

---

[1] Statistical testing can be used, but is not necessary, for metrics using benchmarks.

[2] It would be unreasonable to expect a particular performance metric to always show *ex post* equal or better performance for service to a requesting carrier, compared to that provided to the incumbent LEC's customers. Such a requirement, if implemented, would demand that the incumbent LEC provide *ex ante* superior service to a requesting carrier, in order to ensure that random variation does not cause performance to the requesting carrier to drop accidently below the level needed for a determination of parity.

[3] Other methods of testing are possible, such as the use of Bayesian estimation techniques. We will not discuss those methods here. *See* John Neter, William Wasserman, and G.A. Whitmore, *Applied Statistics* at ch. 27-28 (4th ed., 1993).

[4] Researchers usually call the hypothesis they are trying to prove the alternative hypothesis. The null hypothesis is the hypothesis which they are trying to determine whether to reject. Ramakant Khazanie, *Statistics in a World of Applications* 495 (4th ed., 1997).

hypothesis can be rejected, given the data available.[5] If the data is not consistent with the null hypothesis, then we reject the null hypothesis, and accept the alternative hypothesis.[6] The null hypothesis here would be the hypothesis that the two processes are the same, so that the measurements reflect different observations taken from the same (or identically performing) processes.[7] The alternative hypothesis asserts that the two processes are different.

4. In *Second BellSouth Louisiana Order,* we encouraged BOCs to submit data allowing us to determine if any detected differences in performance are caused by random variation in the data.[8] In its application, Bell Atlantic has presented us with performance data, as well as a statistical test and its corresponding test statistic (called z-scores) that can be used to determine whether a detected difference between the wholesale and retail metrics is statistically significant. Bell Atlantic has been required to utilize this statistical methodology in reporting its performance to New York as part of the Carrier-to-Carrier proceeding.[9]

5. The statistical test that is used depends on the kind of metric being tested, and the number of observations or "sample size" for that metric. The Carrier to Carrier guidelines specify that there are two kinds of metrics, "measured" and "counted."[10] Measured metrics are averages or means of observations (for example, Average Completed Interval).[11] Proportionate (counted) metrics measure the proportion or percentage of a group of observations that meet some criterion (for example, Percentage of Appointments Missed).[12]

---

[5] Devising a statistical test usually involves creating a test statistic and then comparing it to some critical value.

[6] *See* Khazanie, *supra* n.4 at ch. 9.

[7] Statisticians would say that the observations are a sample taken from the population. The population is the theoretical set of values obtained if an infinite number of observations were taken of the underlying process. Therefore the population mean is the theoretical mean produced by the process, while the sample mean is the measured mean. Khazanie, *supra* n.4 at 5-6; Neter, Wasserman, and Whitmore, *supra* n.3 at 235-36, 248-49; Alexander Mood, Franklin Graybill and Duane Boes, *Introduction to the Theory of Statistics* 219-31 (3rd ed., 1974).

[8] *Second BellSouth Louisiana Order,* 13 FCC Rcd at 20659 and n.274.

[9] Bell Atlantic Dowell/Canny Decl. at para. 112, and Attach. B, App. K.

[10] Bell Atlantic Dowell/Canny Decl. Attach. B, App. K. The use of different formulas for statistical testing for measured and proportionate (counted) metrics is recommended in statistical textbooks. *See* Khazanie, *supra* n.4 at 538-48; Neter, Wasserman, and Whitmore, *supra* n.3 at ch. 14. To be consistent with textbook usage, we will refer to "counted" metrics as "proportionate."

[11] Any metric measuring average times is a measured metric. The sample mean, also called the average or the arithmetic mean, is defined as the sum of the observations, divided by the number of observations. Mathematically it is $m = \Sigma X_i / N$, where X_i are the observations, and N is the number of observations. Khazanie, *supra* n.4 at 77-79, 234-35; Neter, Wasserman, and Whitmore, *supra* n.3 at 71, 248-49.

[12] Proportionate metrics are generally said to have a binomial distribution. Neter, Wasserman, and Whitmore, *supra* n.3 at 363-65. Competitive LECs have suggested that there is a third kind of metric involved called rates. Rates are measures that involve the division of two numbers (for example, the trouble rate). Letter from Robert Quinn, Director-Federal Government Affairs, AT&T, to Magalie Roman Salas, Secretary, Federal Communications Commission, CC Docket No. 99-295 Attach. at 13 (Local Competition Users Group, Statistical Tests for Local Service Parity, version 1.0) (filed December 17, 1999) (LCUG Statistical Tests for Local Service Parity). In theory

6. The statistical tests used by Bell Atlantic were initially proposed by Local Competition Users Group (LCUG), a group of competitive LECs.[13] The test LCUG advocated for large sample sizes is commonly known as the "modified z-test", which uses the "modified z statistic."[14] The modified z-test uses only the incumbent LEC's standard deviation, and not the competitive LECs' standard deviation, in calculating the z statistic.[15] It is a variation of the standard textbook z-test, which uses the standard deviations for both the incumbent LEC's and competitive LECs' observations.[16] In its application Bell Atlantic presents us with z-scores,

---

rates can exceed 1, unlike proportions. For example, more than one trouble could be reported for each line, so the trouble rate (which is the number of troubles divided by the number of lines) could be greater than one. Rates are classified as proportionate (counted) metrics by Bell Atlantic, and there are no special formulas for rates used in this application, so we will not discuss formulas for rates here.

[13] The z and t tests to be used for measured and proportionate variables were agreed upon by the Carrier-to-Carrier Group. They were proposed by LCUG and agreed to by Bell Atlantic. LCUG Statistical Tests for Local Service Parity; Bell Atlantic Dowell/Canny Decl. at para. 112, and Attach. B, App. K.

[14] Bell Atlantic Dowell/Canny Decl. Attach. B, App. K. It is also sometimes known as the "LCUG modified z-test." The only known published discussion of the modified z-test is in Cavell Brownie, Dennis D. Boos, and Jacqueline Hughes-Oliver, "Modifying the *t* and ANOVA *F* Tests When Treatment Is Expected to Increase Variability Relative to Controls," *Biometrics* 46, 259-66 (1990). AT&T Pfau/Kalb Aff., Attach. 2, "AT&T's Responses to FCC's Questions Dated April 12, 1999", at 3.

[15] Bell Atlantic Dowell/Canny Decl. Attach. B, App. K. The standard deviation is the square root of the variance. The sample variance is the sum of the squares of the differences between the mean and the observations, divided by the number of observations minus one. For measured metrics the sample variance is: $s^2 = \Sigma (m-X_i)^2 / (N-1)$, where s^2 is the sample variance, m is the sample mean, X_i are the observations, and N is the number of observations. For proportionate metrics the sample variance is: $s^2 = N * P * (1-P)$, where P is the proportion. Khazanie, *supra* n.4 at 257; Neter, Wasserman, and Whitmore, *supra* n.3 at 82-83, 363-71; Mood, Graybill, and Boes, *supra* n.7 at 229.

[16] Assuming the means have *identical* but unknown variances, and the sample size is large, the standard z-test for a difference in means between two populations, stated in terms of competitive LEC and incumbent LEC means, is $z = (m_C - m_I) / (s_p * \text{SQRT}[1/N_C + 1/N_I])$, where m_C = competitive LEC sample mean, m_I = incumbent LEC sample mean, s_p = pooled standard deviation (either uses the observations of both populations, or combines the standard deviations of the two populations), N_C = number of competitive LEC observations, and N_I = number of incumbent LEC observations. The test statistic is normally distributed, so the critical value is obtained from the standard normal distribution. If the sample size is small and the populations are normal, then the standard test is a t-test, using the same test statistic, but the test statistic has a t distribution with $N_I + N_C - 2$ degrees of freedom. Khazanie, *supra* n.4 at 540-41, 563; Neter, Wasserman, and Whitmore, *supra* n.3 at 397-402. The normal and standard normal distributions, t-statistics, and critical values are discussed below at *infra* para. 9 and n.17, 26, 31.

If the variances are assumed to be unknown and *different*, and the sample size is large, then the standard z-test uses the test statistic $z = (m_C - m_I) / \text{SQRT}[s_C^2/N_C + s_I^2/N_I]$, where s_C = competitive LECs' standard deviation and s_I = incumbent LEC's standard deviation. The test statistic is normally distributed, and the standard normal distribution is used to determine the critical value. Khazanie, *supra* n.4 at 538-40, 563. If the sample size is small and the populations are normal, however, then the problem is known as the Behrens-Fisher problem (or the Behrens problem or the Fisher-Behrens problem), which is considerably more complicated to solve. Hamparsum Bozdogan and Donald E. Ramirez, "An Adjusted Likelihood-Ratio Approach to the Behrens-Fisher Problem," *Communications in Statistics: Theory and Methods*, 15 (8) at 2405 (1986); Brownie, Boos, and Hughes-Oliver, *supra* n.14 at 259-60. One solution is to use the Aspin-Welch test, using the same test statistic as for the large sample size test, but here the test statistic has a t distribution, with a complicated calculation of the degrees of freedom. Acheson J. Duncan, *Quality Control and Industrial Statistics* 616-617 (5th ed., 1986); William H. Beyer, *CRC Standard Mathematical Tables* 525 (26th ed., 1987).

which are the test statistic used to perform the z-test.

7. The modified z-test for a difference in means between two populations, assuming the means are normally distributed, used for measured metrics, is:[17]

$$z = (m_C - m_I) / (s_I * \text{SQRT}\ [1/N_C + 1/N_I])$$

where m_C = competitive LEC sample mean, m_I = incumbent LEC sample mean, s_I = incumbent LEC's standard deviation, N_C = number of competitive LEC observations, and N_I = number of incumbent LEC observations. z is the test statistic ("z-score") that results from this calculation.

8. The modified z-test for a difference in proportions between two populations, used for proportionate metrics, is:[18]

$$z = (P_C - P_I) / \text{SQRT}\ [P_I(1-P_I)\ (1/N_C + 1/N_I)]$$

where P_C = competitive LEC sample proportion, P_I = incumbent LEC sample proportion, N_C = number of competitive LEC observations, N_I = number of incumbent LEC observations, and z is the resulting z-score.

9. The z-test involves comparing the z-score for a particular metric with a critical value (call it z_C) to determine if we can reject the (null) hypothesis that the same process generated the Bell Atlantic and competing carrier means. The critical value z_C is chosen based on a particular desired confidence level (call the confidence level C).[19] If the z-score is less than this critical value ($z < z_C$), we reject the null hypothesis, and accept the alternative hypothesis that the processes for serving retail and competing carriers' customers are different. We would then

---

[17] Bell Atlantic Dowell/Canny Decl. Attach. B, App. K. Bell Atlantic calls this a t-test. The t-test uses Student's t distribution to determine the critical value. For large sample sizes, this is approximately equivalent to doing a z-test, because for sample sizes of greater than 30 observations, the t distribution has approximately the same distribution as the standard normal distribution. With a t-test, the critical value varies according to the degrees of freedom, which depend on the sample size. Since Bell Atlantic is using a fixed critical value, it is effectively using a z-test. The formula for calculating the test statistic is effectively the same for both kinds of tests. Khazanie, *supra* n.4 at 410-413, 521; Neter, Wasserman, and Whitmore, *supra* n.3 at 913; *see infra* n.31. The formula for the test statistic can be more simply described as the difference in means divided by the standard error (Bell Atlantic calls the standard error the "sampling error" in the Carrier to Carrier metric reports), or $(m_C - m_I)$ / S.E. Neter, Wasserman, and Whitmore, *supra* n.3 at 266, 290-91; LCUG Statistical Tests for Local Service Parity at 7-8.

[18] Bell Atlantic Dowell/Canny Decl. Attach. B, App. K. We note that the standard z-test for a difference in proportions is $z = (P_C - P_I) / \text{SQRT}\ [P_p(1-P_p)\ (1/N_C + 1/N_I)]$, where P_C = competitive LEC sample proportion, P_I = incumbent LEC sample proportion, P_P = pooled sample proportion, N_C=number of competitive LEC observations, and N_I=number of incumbent LEC observations. Khazanie, *supra* n.4 at 546-47, 563; Neter, Wasserman, and Whitmore, *supra* n.3 at 408-12.

[19] As noted above, the critical value for a z-test is taken from tables based on the standard normal distribution. *See supra* n.16.

say that the test indicates the measured difference in metric values is statistically significant.[20] If the confidence level is C, then the probability of mistakenly rejecting the null hypothesis when it is true would be 1-C (call this α).[21] Statisticians call α, the probability of mistakenly rejecting the null hypothesis, the probability of a Type I error.[22] The confidence level can be interpreted as our confidence that we have not mistakenly rejected the null hypothesis (i.e., found a difference to be statistically significant when it is not).[23] Thus if we use the 95 percent confidence level for a one-tailed test,[24] the critical value (taken from tables) is -1.645, and there is a 5 percent probability that a statistically significant difference will be detected when the process in fact is the same.[25]

10. Z-tests, including the modified z-test and the standard z-test, are only appropriate if the distribution of the mean (or of the proportion, for proportionate measures) is normal.[26] Even for metrics whose observations are not normally distributed, the mean should be normally distributed if the sample size is large enough, according to the Central Limit Theorem.[27] Usually it is assumed that a sample size of 30 or more is sufficient for it to be appropriate to use the z-test

[20] In New York the tests have been set up so that z-scores that indicate worse performance for competing carriers are negative. Thus if the critical value is –1.645, only z-scores that are less, such as –2 or –3, would yield statistically significant results.

[21] For example, if the confidence level C is 95 percent (0.95), then α is 5 percent (0.05).

[22] While falsely rejecting the null hypothesis (i.e., falsely finding that the BOC's processes of serving retail and competitors' customers are different) is called a Type I error, falsely accepting the null hypothesis when it is not true (i.e., falsely finding that the BOC's processes are identical) is called a Type II error. The probabilities of a Type I error and a Type II error are commonly referred to by the Greek letters α (alpha) and β (beta), respectively. Khazanie, *supra* n.4 at 498; Neter, Wasserman, and Whitmore, *supra* n.3 at 319-20. Usually statisticians choose one hypothesis to be the null hypothesis because falsely rejecting it (Type I error) is considered more serious than falsely accepting it (Type II error), so controlling α is more important than controlling β. Khazanie, *supra* n.4 at 499, 506; Neter, Wasserman, and Whitmore, *supra* n.3 at 320; Mood, Graybill, and Boes, *supra* n.7 at 411.

[23] Statistical tests virtually never determine anything with certainty. There is always a certain probability of being wrong and choosing the incorrect hypothesis. Statistical tests are devised to minimize this probability of being wrong, i.e., to keep the probabilities of Type I and Type II errors at a minimum.

[24] The rationale for using a one-tailed test is described below. *See infra* para. 18.

[25] This means that, if a 95 percent confidence level is used for a statistical test, when the null hypothesis is true, 95 percent of the time we will correctly choose the null hypothesis. Meanwhile there will be a 5 percent chance that a statistical test will show a statistically significant difference. This is caused by random variation in the data. One way to interpret this is that out of every 100 measurements, on average five should show statistically significant differences, even with identical processes serving retail and competing LECs' customers.

[26] A normal distribution is sometimes referred to as a Gaussian distribution. It is often described as having a "bell-shaped" curve. A standard normal distribution is a normal distribution that has been transformed such that its mean is zero and standard deviation is one. Khazanie, *supra* n.4 at 281, 294-96.

[27] The Central Limit Theorem is a powerful theorem in statistics. It says that under most circumstances, the distribution of the mean will approach a normal distribution for a large enough sample size, even if the distribution of the population from which the mean is drawn is not normal. Khazanie, *supra* n.4 at 344-45; Neter, Wasserman, and Whitmore, *supra* n.3 at 267-68; Mood, Graybill, and Boes, *supra* n.7 at 233-36.

for measured metrics.[28] For proportionate metrics, it is generally assumed that a z-test can be used if the sample size is large enough such that $N * P \geq 5$ or $N * (1-P) \geq 5$.[29]

11. For metrics with small sample sizes, Bell Atlantic is using the binomial test, t-test, and the permutation test. For proportionate measures with small sample sizes, defined as $N * P * (1-P) < 5$, where N is the number of observations and P is the proportion, Bell Atlantic will use a binomial test to test whether the difference in proportions is statistically significant.[30] For measured metrics with small sample sizes (less than 30 observations), Bell Atlantic is temporarily using a t-test, which assumes the population is normally distributed, or close to a normal distribution.[31] However, a non-parametric test should be used if the population is not normally distributed. Non-parametric tests do not assume the data or the mean have a particular distribution. Bell Atlantic is committed to using a permutation test, which is one kind of non-parametric test, to determine if differences in performance between Bell Atlantic retail customers and competitive LECs are statistically significant, once it is able to implement it for all metrics.[32]

12. Unlike standard z-tests, the modified z-test assumes that the incumbent LEC and competitive LEC variances are the same under parity (the null hypothesis), but not necessarily so

---

[28] Textbooks are vague about what the minimum sample size should be to use a large sample test like the z-test on measured metrics, but 30 is often cited as appropriate. Textbooks generally agree, however, that at 30 observations and greater, the t-test can be replaced by the z-test, for distributions that are approximately normal. *See, e.g.*, Khazanie, *supra* n.4 at 413, 521, 539; Neter, Wasserman, and Whitmore, *supra* n.3 at 913; Duncan, *supra* n.16 at 150. *See supra* n.17. Doubts about whether a sample size of 30 is sufficient for measured metrics have been raised by AT&T in other proceedings. AT&T Pfau/Kalb Aff. Attach. 3, at 4-5. We note that KPMG used 100 as the threshold for using permutation testing in their test analysis of Bell Atlantic's metrics. KPMG Final Report at POP8 IV-176-77. The parties in this proceeding have agreed to use 30 as the minimum sample size for use of a z-test, and there is insufficient evidence in the record for us to reject this choice. Bell Atlantic Dowell/Canny Decl. Attach. B, App. K. The minimum sample size needed before a z-test should be used generally depends on the distribution of the underlying observations, and, in particular, how skewed it is. Neter, Wasserman, and Whitmore, *supra* n.3 at 296.

[29] Khazanie, *supra* n.4 at 262-64; Neter, Wasserman, and Whitmore, *supra* n.3 at 368-69.

[30] Bell Atlantic Dowell/Canny Decl. Attach. B, App. K. KPMG used a hypergeometric test (also known as Fisher's Exact Test) for its analysis when the number of observations is less than 10,000, for comparing two proportions. KPMG Final Report at POP8 IV-177. For a discussion of the binomial and hypergeometric distributions and tests, which are similar, *see* Khazanie, *supra* n.4 at 246-64; Neter, Wasserman, and Whitmore, *supra* n.3 at ch. 7.

[31] The t-test is similar to a z-test. Unlike a z-test, it is used for small sample sizes, when the population is assumed to be normal, and the variance is not known. The t-test uses the same formula for the test statistic as the z-test (*see supra* para. 7), but instead of obtaining the appropriate critical value from a table of the standard normal distribution, the critical value has to be taken from the tables for the t distribution, taking into account the appropriate degrees of freedom (i.e., number of observations). Note that the t-test yields about the same result as a z-test for sample sizes of 30 or more. Since z-tests are easier to do, they are usually used for large sample sizes. *See supra* n.17; Khazanie, *supra* n.4 at 410-413, 521; Neter, Wasserman, and Whitmore, *supra* n.3 at 335-36, 402-03, 913.

[32] Bell Atlantic says it will initially use a t-test until it is able to run a permutation test in "an automated fashion." Bell Atlantic Dowell/Canny Decl. Attach. B, App. K.

under the alternative hypothesis.[33] With this test, unlike a standard z-test, z-scores will not fall if competitive LECs' standard deviations rise.[34] While it is a test of a difference of means, it will also be more likely to show a statistically significant difference if the competitive LEC variance is larger.[35] This means this will also serve as a weak test for a difference of variances.

13. We find the modified z-test, the binomial test, the t-test, and the permutation test to be reasonable tests for statistical significance, for measured and proportionate measures. All parties in the New York Commission collaborative hearings have agreed to the use of these tests, and these tests have been adopted for use in the Carrier-to-Carrier measures and the Performance Assurance Plan.[36] Moreover, no commenters in this proceeding have objected to the use of the modified z-test, the t-test, the binomial test, or the permutation test. These tests are efficient in their ability to detect differences in means or proportions that are not caused by random fluctuation, while minimizing the likelihood of falsely concluding the variation may be due to underlying discrimination. They appear to be relatively powerful tests.[37] We find the modified z-test (t-test for small sample sizes) to be a reasonably efficient test to determine whether a difference in means or proportions is statistically significant. We further find that the two nonparametric tests proposed, the binomial and the permutation tests, are both fairly standard tests to use when the samples are small. The permutation test is a standard nonparametric test

---

[33] In other words, it assumes: H_0: $\mu_I = \mu_C$ and $\sigma_I^2=\sigma_C^2$, and H_A: $\mu_I \neq \mu_C$ or $\sigma_I^2\neq\sigma_C^2$, where μ is the population (theoretical) mean for the incumbent I and competitive LEC C, σ^2 is the variance, and H_0 and H_A are the null and alternative hypotheses, respectively. Brownie, Boos, and Hughes-Oliver, *supra* n.14 at 260; LCUG Statistical Tests for Local Service Parity at 8-9. There are no standard textbook z-tests for these hypotheses. There are standard z-tests for a test of difference of means which assume that the incumbent and competitive LEC variances are always the same, or that the variances are always different. *See* Khazanie, *supra* n.4 at 563; Neter, Wasserman, and Whitmore, *supra* n.3 at 538-42, 563; *supra* n.16.

[34] In a standard z-test, if the competitive LECs' standard deviation rises, so will the standard error (the denominator in the z statistic), causing the z statistic to fall, even if the difference in the means stays constant. This will not happen with the modified z, since its standard error does not directly depend on the competitive LECs' standard deviation.

[35] If the competitive LEC variance (and standard deviation) is large, then the competitive LEC means m_C will be much more variable. Since the standard error for the modified z does not depend on the competitive LEC standard deviation, unlike the standard z, the modified z will be more likely to find that a difference in means is statistically significant. AT&T Pfau/Kalb Aff., Attach. 2, "AT&T's Responses to FCC's Questions Dated April 12, 1999" at 3-4.

[36] Bell Atlantic Dowell/Canny Decl. at para. 112, and Attach. B, App. K, and Attach. C, Ex. 1, App. D; AT&T Pfau/Kalb Aff. at para. 54.

[37] Statisticians define the power of a test as its ability to correctly determine when the alternative hypothesis is true, while keeping fixed the probability of falsely rejecting the null hypothesis, for every possible alternative hypothesis (or Power=1-β while α is fixed, for all H_A). A more powerful test has a lower β, for the same α and H_A. Neter, Wasserman, and Whitmore, *supra* n.3 at 339-47; Mood, Graybill, and Boes, *supra* n.7 at 406-11; William H. Greene, *Econometric Analysis* 156-57 (3rd ed., 1997). Therefore, these tests are more powerful if they are better able to detect differences in means when the processes serving retail customers and competitors are truly different, while maintaining the same probability of falsely finding a difference when the processes are, in fact, the same. The modified z has been shown to be a more powerful test than a standard z under the hypotheses outlined above (*supra* n.33), using power curves. Brownie, Boos, and Hughes-Oliver, *supra* n.14 at 261-63.

used to test for a difference in means for small samples.[38] We note that the binomial test is considered to be an exact test for proportionate metrics, such that it is the most powerful test possible.[39]

14. We will rely on the results of the tests and their associated test statistics that Bell Atlantic has presented to us with this application. However, we do not rule out the use in other section 271 applications of alternative statistical tests that are of similar power and efficiency. For measures where the New York Commission has identified retail analogues, we will use the modified z-scores presented by Bell Atlantic to determine if a difference in performance provided to competitive LECs' and Bell Atlantic's retail customers is statistically significant. As discussed below, we will employ a 95 percent confidence level one-tailed test, which yields a critical value (or minimum threshold z-score) of –1.645.[40] We note that the New York Commission has adopted this confidence level and critical value for its determination of performance scores of –2 for the Performance Assurance Plan.[41]

15. Therefore we will treat all z-scores that are positive, or are larger than -1.645, as evidence of nondiscrimination.[42] Positive z-scores indicate that competitive LEC customers received better performance than Bell Atlantic retail customers. Z-scores between zero and -1.645, such as a score of -1, indicate that competitive LECs received on average poorer service than Bell Atlantic retail customers, but that there is a significant likelihood that Bell Atlantic's process of serving both sets of customers was identical, and the negative score was due to random chance. In these cases the difference would not be considered statistically significant, and we would conclude that Bell Atlantic has met its burden of demonstrating nondiscrimination. Z-scores of less than -1.645, such as a score of -2 or of -3, would be viewed as statistically significant. Only in the last case would we then conduct a further inquiry into whether the difference is large enough to be deemed discriminatory.

16. The Carrier-to-Carrier guidelines have set no minimum sample size, so that statistical tests are reported even if the sample size is just one observation.[43] We make no determination here as to whether it is reasonable to have a minimum sample size for statistical

[38] Permutation tests are classified as a bootstrap method. Bootstrap methods involve repeated resampling of the original data to generate the statistical results of interest. A.C. Davison and D.V. Hinkley, *Bootstrap Methods and Their Applications* at chs. 1, 4 (1997); H. Scheffe, *The Analysis of Variance* 313-18 (1959).

[39] Duncan, *supra* n.16 at 608, 973-75.

[40] *See infra* para. 17. The Carrier-to-Carrier metrics are set up in such a way that negative scores indicate that competitive LECs are receiving worse performance than Bell Atlantic customers, while positive scores indicate the opposite. *See* Bell Atlantic Dowell/Canny Decl. Attach. C, App. D at 1.

[41] The plan also provides for performance scores of –1, which represent a confidence level of 79 percent. The adjustment used in the plan of erasing a –1 if followed by zeros in two following months effectively raises the confidence level to 90 percent for –1's that are not erased. Bell Atlantic Dowell/Canny Decl. at paras. 128-29, and Attach. C, App. E at 1.

[42] Note that a "larger" negative score is actually closer to zero, so –1 is larger than –2.

[43] Bell Atlantic Dowell/Canny Decl. Attach. B, App. K.

testing. We believe, however, that the data should be reported for all sample sizes, so that we will have some information about performance for all services provided.[44] We note that for some kinds of orders, such as those for collocations or for high capacity lines like DS3s, small numbers of observations are possible for a given month. The importance and large revenues involved for each observation makes it important for us to have information about these orders.

17. When we look at the differences in metric values, we will assume that parity exists unless the competitive LEC scores are worse than those for the BOC, and the difference is statistically significant at the 95 percent confidence level for a one-tailed test.[45] We use the 95 percent confidence level because it is a commonly used standard, and because it gives us a reasonable likelihood of detecting variations in performance not due to random chance, with few false conclusions that variations are not due to random chance.[46] At the 95 percent confidence level, even under parity an average of 5 percent of the tests should fail (this is the probability of a Type I error).[47] At higher confidence levels this probability would be lower, but then the probability of not detecting unexplained variations in performance if they do exist (the probability of a Type II error) would increase. The 95 percent confidence level appears to be a fair compromise. We do not comment here on AT&T's proposal to choose a confidence level of 85 percent, which it says will balance the probability of Type I and Type II errors.[48] We find that AT&T has not put sufficient evidence on the record for us to determine that setting the confidence level at 85 percent[49] will in fact balance the probability of Type I and Type II errors.[50]

---

[44] For metrics with observations excluded from their measurement, the number of observations excluded should also be reported, to improve our ability to determine how accurately the metric measures the universe of orders or customers.

[45] A difference in metric values that is statistically significant, however, does not necessarily mean that the BOC's service is discriminatory. We will examine the totality of the evidence before making a determination whether the BOC is providing parity.

[46] Khazanie, *supra* n.4 at 506; Neter, Wasserman, and Whitmore, *supra* n.3 at 298. We note that Bell Atlantic argues that the 95 percent confidence level is appropriate. Bell Atlantic Dowell/Canny Decl. Attach. B, App. K; Bell Atlantic Duncan Reply at para. 36-38.

[47] Type I and Type II errors are described above. *See supra* para. 9.

[48] AT&T argues that choosing a critical value to balance the probabilities of Type I and Type II errors is desirable, because it balances the interests of BOC and competitive LECs by setting equal the chances of falsely finding discrimination and of falsely missing discrimination. While acknowledging that the critical value to achieve this balancing ("balancing critical value") will depend on the number of BOC and competitive LEC observations, they argue that using a fixed critical value based on an 85 percent confidence level is a reasonable approximation of the balancing critical value, given typical competitive LEC sample sizes. AT&T Pfau/Kalb Aff. at paras. 88-93 and n.97 and Attach. 2 at 27-30.

[49] This would mean using a critical value for the z-test of 1.04.

[50] AT&T's proposal to balance the Type I and Type II error probabilities does appear to have the attractive feature that the interests of the incumbent LEC and the competitive LECs are given equal weight, so that the probabilities of falsely concluding the incumbent LEC may be discriminating and of missing existing discrimination are balanced (so $\alpha=\beta$). Such an approach could be used in future section 271 applications. We would be more likely to accept use of such an approach if the state commission and parties have agreed on its use, particularly since there are

18. We accept Bell Atlantic's use of a one-tailed statistical test. We find a one-tailed test appropriate because we are only concerned with *inferior* performance provided by the incumbent LEC to the competitive LEC. Therefore we are only testing to determine whether inferior performance that is being provided by the incumbent LEC to a competitive LEC is statistically significant.[51] We note that the New York Commission has approved the use of a one-tailed test, and no commenters object to its use.[52]

19. For metrics that have no retail analogue, Bell Atlantic presents us with a benchmark level adopted by the New York Commission, and no statistical comparison is employed. According to the Carrier to Carrier guidelines, Bell Atlantic would fail a benchmark test if performance to competing carriers falls below the benchmark level.[53] We accept Bell Atlantic's use of benchmarks without a statistical test being employed. We make no determination here whether it would be better to employ a statistical test or a straight comparison. We accept, however, the use of a direct comparison, which we are presented with here.

---

details that need to be worked out before it is used. For example, the relevant alternative hypothesis must be agreed upon. We note that the New York Commission has not accepted AT&T's proposal. Bell Atlantic argues that AT&T's proposal is not standard and is difficult to implement. Bell Atlantic Duncan Reply at paras. 36-38.

[51] The alternative is to use a two-tailed test to determine whether an incumbent LEC's performance to competitive LECs is either inferior *or superior* to the performance that it provides itself. Our analysis does not take into account whether superior performance is being provided. We are unable to determine how much superior performance in one metric or for one month could offset inferior performance in another metric or for another month.

[52] Bell Atlantic Dowell/Canny Decl. Attach. B, App. K. The use of passing scores in some months to offset negative scores in other months is used in the Performance Assurance Plan to lower the probability of Bell Atlantic making payments under parity. Bell Atlantic Dowell/Canny Decl. at paras. 128-29. *See supra* n.41. This is one reasonable method of reducing the probability of a Type I error.

[53] Bell Atlantic Dowell/Canny Decl. Attach. C, Ex. 1 at 4 and App. C. *See supra* Section III.C.2.

APPENDIX C: ANALYSIS OF AVERAGE COMPLETED INTERVALS FOR NON-DISPATCH ORDERS USING CARRIER TO CARRIER AND GERTNER/BAMBERGER STUDY DATA

1. In this appendix we adjust the reported Average Completed Interval data for competing carriers' orders to correct for the factors Bell Atlantic cites. In this manner, we can make a proper comparison of the Bell Atlantic retail and competing carrier intervals. According to Bell Atlantic, the disparity between retail and wholesale Average Completed Intervals for non-dispatch orders is due to two factors: (1) the improper coding by competing carriers of some "W" coded orders, when they request longer intervals than the standard interval; and (2), competing carriers' customers requesting a mix of services that have longer standard intervals associated with them, compared to the mix of services requested by Bell Atlantic's retail customers.[1] Using the Gertner/Bamberger study's results, it is possible to see whether correcting for these factors would explain the evident difference between Bell Atlantic retail and wholesale Average Completed Intervals in the reported Carrier to Carrier metrics for non-dispatch orders.[2] As set forth below, we find that, after accounting for those factors, a half day difference between wholesale and retail Average Completed Intervals remains for UNE-P orders, and for resale orders, a quarter day difference remains for July and August, while the intervals are about equal in June.

a. Analysis of UNE-P Orders

2. We make the following calculations. The data in the Gertner/Bamberger study allows us to estimate the Average Completed Interval for competing carriers' properly coded "W" orders, and make an adjustment for the differences in order mix. The calculations we make, and the resulting differences that we find for non-dispatch UNE-P orders (measured in days), are summarized in the Table below.

[1] Bell Atlantic Dowell/Canny Decl. at paras. 62, 65-66; Bell Atlantic Gertner/Bamberger Decl. at paras. 7-12.

[2] The Gertner/Bamberger study provides us with no information about the impact of the factors they discuss on dispatch orders, so we are unable to make the same adjustments for Average Completed Intervals for dispatch orders.

Estimated Difference in Average Completed Intervals for Non-dispatch UNE-P Orders[3]

	June			July			August		
	BA	CLEC	Diff	BA	CLEC	Diff	BA	CLEC	Diff
Carrier to carrier metrics data	1.25	3.20	-1.95*	0.99	2.55	-1.56*	1.07	1.91	-0.84*
Using properly "W" coded CLEC orders	1.25	1.13	0.12	0.99	1.31	-0.32	1.07	2.36	-1.29
Adjustment to CLEC data for difference in standard intervals		+0.53			+0.04			-0.62	
CLEC data revised for alleged biases	1.25	1.68	-0.43*	0.99	1.35	-0.36*	1.07	1.74	-0.67*

3. The top line in the table is the Average Completed Interval data reported in the Carrier to Carrier report for both Bell Atlantic retail orders and competing carriers' ("CLEC") orders, which Bell Atlantic claims is flawed because of improper "W" coding and the order mix problem. The second line compares the Bell Atlantic retail interval from the Carrier to Carrier report with the Average Completed Interval data from the study for properly "W" coded competing carriers' orders. The third line shows the adjustment made to the competing carriers' measured intervals to account for differences in the average standard intervals, caused by the order mix problem. The bottom line compares the adjusted competing carriers' data, which has been corrected for the "W" coding and order mix problems, with the Bell Atlantic retail data. The table shows that the Average Completed Interval for competing carriers is much smaller after these corrections are made for the "W" coding and order mix problems. Specifically, the difference between Bell Atlantic retail and competing carriers' orders is about half a day, and is statistically significant.[4]

[3] Sources are Carrier to Carrier metrics, Bell Atlantic Dowell/Canny Decl. Attach. D; Bell Atlantic Gertner/Bamberger Decl. at Table 4; Bell Atlantic Gertner/Bamberger Reply Decl. at Table 2. The Bell Atlantic retail numbers used for comparison with the study data for CLECs were taken from the carrier to carrier metrics. The bottom row includes an adjustment to the CLEC average completed interval to take into account the different lengths of the average standard intervals (listed in the third row). The calculation of the CLEC intervals in the bottom row involved taking the study's estimate of the interval for only properly coded orders from Table 4 (2.36 days in August) and adding the difference in average standard intervals between retail and CLEC orders caused by the different order mixes, taken from Table 2 of the Reply (1.84–1.22=0.62 days in August), to get the revised CLEC interval (2.36–0.62=1.74). The column "Diff" contains the differences between Bell Atlantic and CLEC intervals. Results that appear to be statistically significant are marked with an asterisk. *See infra* n.8.

[4] Statistical significance is determined by calculating a z-score, which is the difference in the means divided by the standard error (called "sampling error" by Bell Atlantic), and then examining whether the z-score is less than –1.645. In order to determine whether our estimated differences in Average Completed Intervals are statistically significant, the standard error must be recalculated. The standard error used here differs from the value published in the Carrier to Carrier report because the number of CLEC orders in that report was used in its calculation, and that number was inflated because of the number of miscoded orders included in it. The standard error is: $SE = s_I$ SQRT $[1/N_C+1/N_I]$, where s_I is the standard deviation for Bell Atlantic, N_C is the number of CLEC observations, and N_I is the number of observations for Bell Atlantic. For our calculations s_I and N_I are the same as in the Carrier to Carrier (continued....)

b. Analysis of Resale Orders

4. Although the Carrier to Carrier data is disaggregated between business and residential orders, the Gertner/Bamberger study data is not. In order to perform our analysis, we aggregated the business and residential Carrier to Carrier data. We then used the data from the Gertner/Bamberger study to estimate the Average Completed Interval for competing carriers' properly coded "W" orders, and make an adjustment for the differences in order mix, as we did above for UNE-P orders. The calculations we make to the competing carriers data, and the resulting differences that we find for non-dispatch resale orders (measured in days), are summarized in the Table below.

Estimated Difference in Average Completed Intervals for Non-dispatch Resale Orders[5]

	June			July			August		
	BA	CLEC	Diff	BA	CLEC	Diff	BA	CLEC	Diff
Carrier to carrier metrics data[6]	0.96	1.90	-0.94*	1.01	1.59	-0.58*	1.06	1.58	-0.52*
Using properly "W" coded CLEC orders	0.96	0.86	0.10	1.01	1.10	-0.09	1.06	1.15	-0.09
Adjustment to CLEC data for difference in standard intervals		+0.07			+0.19			+0.10	
CLEC data revised for alleged biases	0.96	0.93	0.03	1.01	1.29	-0.28*	1.06	1.25	-0.19*

5. As evidenced by the bottom line of this table, the differences in Average Completed Intervals for resale orders between competing carriers and Bell Atlantic's retail customers are much smaller than before the correction. In fact, the Average Completed Intervals

(Continued from previous page)
report. We adjust the published N_C to remove miscoded orders from the count. This was done using the percentage of orders that were miscoded, which was provided in the right column of Table 1, in Bell Atlantic Gertner/Bamberger Reply. So, for example, in August for UNE-P there were 25,270 Bell Atlantic orders (N_I) and Bell Atlantic's standard deviation was 2.35 (s_I). There were 10,642 CLEC orders, of which 45.9 percent were miscoded, leaving 5,757 orders correctly coded (N_C). The result is a standard error of 0.034. The calculated standard error for July was 0.034, and for June was 0.043. The resulting z-scores are –10.1, –10.7 and –19.5, for June, July, and August, all of which are statistically significant.

[5] Sources are Carrier to Carrier metrics, Bell Atlantic Dowell/Canny Decl. Attach. D; Bell Atlantic Gertner/Bamberger Decl. at Table 4; Bell Atlantic Gertner/Bamberger Reply Decl. at Table 2. The Bell Atlantic retail numbers were aggregated from the Carrier to Carrier metric data on business and residential orders, to allow comparison with the study's numbers for CLECs. For the calculations of the adjusted CLEC numbers. *See supra* n. 4. Results that appear to be statistically significant are marked with an asterisk. *See infra* n.8.

[6] Both retail and CLEC data are aggregated for both business and residential orders.

are about equal in June for wholesale and retail orders.[7] In July and August, the differences are about a quarter day, but are, nevertheless, statistically significant.[8]

---

7 The measured difference of 0.03 days is not likely to be statistically significantly different from zero. *See infra* n.8.

8 Calculations of statistical significance were made using the same formulas as in n.4, except that calculating an aggregate standard error was more difficult for resale orders because the standard deviations are provided only in disaggregated form, for business and residential orders. The business and residential numbers of observations were added to yield the total numbers of observations for Bell Atlantic (N_I) and CLECs (N_C). The aggregate Bell Atlantic standard deviation was approximated by taking the weighted average of the business and residential standard deviations, weighted by the number of observations. This should yield a standard deviation close to the true standard deviation for the pooled set of observations, if the means for business and residential customers are close together. The means are close for August (1.07 for business versus 1.06 for residential) and for July (0.99 for business versus 1.01 for residential), and the calculated standard error for August is 0.049, and for July is 0.044. The means are not close for June (1.25 for business versus 0.94 for residential), but the Average Completed Intervals show that competing carriers received better service than retail customers in June. The calculated z-scores are –3.9 for August, and –6.3 for July, both of which are statistically significant. If we use July or August's standard errors, it is apparent the June difference of +0.03 days is not statistically significantly different from zero.